



5th Annual Rocky Mountain  
Bioinformatics Conference  
November 30 - December 2, 2007  
Aspen/Snowmass, Colorado Silvertree Hotel



## POSTER ABSTRACTS – *updated 12/07/07*

---

### *DNA-binding Affinity Scale for Amino Acids from Experimental Binding Data*

**Presenter:** Shandar Ahmad

**Author(s):** Shandar Ahmad

**Abstract:** Base-amino acid interactions are the basic unit in the recognition of DNA by proteins. Large amount of thermodynamic data relating to mutations in DNA-binding proteins is now available. In this work, an analysis of this data is presented. A linear scale to estimate the contribution of individual amino acid residues to binding free energy in protein-DNA interactions was developed. This scale seeks to maximize the accuracy of predicting free energy change in single substitutions (ddG) in experimental binding data of mutants obtained from public sources. Based on this scale 69.0 to 71.4% of reported amino-acid mutations could be correctly classified between those stabilizing and destabilizing protein-DNA interaction, using a five-fold cross-validation. Estimates of free energy change were made at a mean absolute error (MAE) of 0.89 kcal/mol and Pearson's correlation of 0.411 (in a sigmoidal space, used for prediction, an average correlation of 0.496 was observed). In this scale, basic residues-as expected- have a high affinity score (Arg: 6.1, Lys: 7.0), whereas acidic residues (Asp: -7.02, Glu: -5.68) and the smallest residue Gly (-7.73) got the lowest scores for DNA-binding affinity. Resulting affinity values (F-scores) have been compared with known properties of amino acids and best correlated properties- charge, helix capping preferences and hydrophobicity- have been discussed.

---

### *Base-specific Prediction of DNA-binding Sites in Proteins*

**Presenter:** Munazah Andrabi

**Author(s):** Munazah Andrabi, Shandar Ahmad, Kenji Mizuguchi, Akinori Sarai

**Abstract:** There has been an extensive interest in the study of DNA-binding proteins in the recent past. Some studies have focussed on predicting transcription factor binding sites on DNA, whereas there are others which focussed on predicting if a new protein could be a DNA-binding one. Earlier, we have thoroughly analyzed the sequence and structural features of DNA-binding sites in proteins and developed a novel method of predicting residues participating in these interactions using neural network [Ahmad et al 2004, Ahmad and Sarai 2005]. Similar studies on DNA-binding prediction have appeared since. In this work, we have attempted to predict DNA-binding residues, specific to each base. Thus, we have created data sets of proteins with residue-wise information of their binding to each of the four bases, calculated the propensity scores for each combination and developed neural network models for each one of them. Results of prediction show subtle differences between the propensity scores and base specificity of binding sites. We have also benchmarked prediction performance of publicly available methods of

predicting DNA-binding residues on a non-redundant data set derived from protein-DNA complexes submitted to PDB after the publication of these web servers. We find that models trained on unusually larger windows of information and using redundancy in data sets show exaggerated performance during training which could not be sustained on new proteins submitted after these publications.

---

***Computational Examination of Gain and Loss of Phosphorylation Consensus Sites in Cancer***

**Presenter:** Peter Baenziger

**Author(s):** Peter Baenziger, Uday Shanker, Matthew Mort, Sean Mooney, Maricel Kann, Matthew Hahn, Predrag Radivojac

**Abstract:** Many specific mutations have been studied and we now understand the basic biochemical functions these mutations disrupt to cause a clinical observation of disease. However, both the Human Gene Mutation Database and NCBI's dbSNP contain tens of thousands of amino acid substitution causing mutations of which most are not annotated with phenotype or underlying biochemical effect (HGMD: 60,000; dbSNP: 40,000). Understanding how a mutation causes biochemical changes that lead to a disease is a formidable problem which will require years to unravel. Here we study the relationship between phosphorylation site disruption and disease using predictions of phosphorylation on sites mutated in cancer. In contrast to other studies of phosphorylation and disease, we are studying changes in phosphorylation target sites, as opposed to kinase or phosphatase dysregulation. Our data sets include two sets of somatic cancer mutations, disease associated mutations from Swiss-Prot, evolutionary mutations in orthologous sequences, nonsynonymous polymorphisms from two populations in gene resequencing projects and random amino acid substitutions derived from these protein sets. Our work shows cancer data sets are statistically enriched in gain or loss of a phosphorylation site through mutation when compared to polymorphic positions in populations and orthologous positions in other organisms. This evidence suggests phosphorylation may be an important feature for predicting disease causing mutations and could represent a molecular cause of disease. Interestingly, we also find kinases represent the most enriched set of phosphorylation site disruption, suggesting phosphorylation target site mutation is an active cause of phosphorylation dysregulation.

---

***Deciphering Transmembrane Helix Recognition by the ER Translocon***

**Presenter:** Andreas Bernsel

**Author(s):** Andreas Bernsel

**Abstract:** Recognition of transmembrane (TM) alpha-helices in integral membrane proteins is carried out by the Sec61 translocon in the endoplasmic reticulum membrane. The physicochemical properties of these lipid bilayers are highly dependent on the distance to the membrane center, and this is reflected in the variable amino acid composition of TM helices. Here, we develop a quantitative description of the contribution to membrane insertion efficiency for all twenty amino acids as a function of

position within the helix, based on experimental data from in vitro translation and membrane integration of a large number of systematically designed hydrophobic segments. In addition, the effects from TM segment length and flanking amino acids are analyzed. The position-specific scale of amino acid contributions resulting from this analysis is implemented in a simple hidden Markov model-based method to predict TM topology. In a benchmark on known 3D structures of membrane proteins, our simple method is found to perform on par with the current best statistics based TM topology predictors, which often contain hundreds of parameters optimized on known membrane protein topologies. TM helices containing e.g. charged residues towards the interfacial region that are missed by other methods, can typically be found using the position specific information inherent in our method.

---

### ***Systematic Order Dependent Effect in Affymetrix GeneChips***

**Presenter:** Kathe Bjork

**Author(s):** Kathe Bjork, Karen Kafadar

**Abstract:** The Affymetrix GeneChip 3' expression profiling system contains an order dependent (OD) effect that creates bias in transcript abundance measurements with the subsequent bias carried forward into inference derived from chips. OD has not been previously reported in Affymetrix corporate literature, and was only recently reported in bioinformatics literature, although it is present in three generations of human GeneChips. It is easily visualized by plotting perfect match and mismatch values, and summary expressions such as MAS5.0 signal, PLIER, robust multi-array average (RMA) and GC-RMA values, in alphanumerically ascending order, the default output from Affymetrix GeneChip and Bioconductor 'affy' package algorithms. Using publicly available HG-U133Plus2.0, HG-U133A, HG-U95Av2, Mouse 430A, and Mouse 74Av2 GeneChip profiles, we describe the characteristics of OD and its influence on inference, including bias in detection calls, variance and differential expression. We also describe current transformations and models under investigation for its remediation.

---

### ***Selecting a Scoring Matrix***

**Presenter:** Promita Bose

**Author(s):** Promita Bose, Robert Edwards

**Abstract:** Substitution matrices are among the most widely used scoring techniques throughout bioinformatics and information theory: BLAST, PHYLIP and other alignment packages, all use them. These widely used matrices ignore organism specific properties and do not provide customized scoring schemes for groups of organisms. We have developed organism-specific scoring matrices for phage genomes. Phages are viruses that specifically infect bacteria, and are not subject to the normal constraints of whole genome evolution. Three different models of amino acid replacement were used to build the substitution matrices. These models use information from approximately five and a half million protein alignments from over five hundred phage genomes. The models are shown to be specific for phages and suggest unique aspects of phage evolution and

biology. The scoring matrices are different from the PAM and BLOSUM series of matrices and demonstrate the need for specific matrices from other groups, or clades, of organisms.

---

***Advancing X-ray Crystallography via Protein Structure Prediction***

**Presenter:** Christopher Bottoms

**Author(s):** Christopher Bottoms, Jingfen Zhang, Rajkumar Bondugula, John Tanner, Dong Xu

**Abstract:** By knowing the structures of proteins, we can design drugs that target specific diseases. If a protein can be crystallized, then X-rays can be used to get information from the protein crystal that can lead to solving its structure. Currently, this process requires collecting only one data set if a structure of a very similar protein (i.e. at least 30% identical) is known. This technique, well-known among protein crystallographers, is called molecular replacement. The known structure serves as a molecular replacement model for the unknown structure. For many proteins, a suitable molecular replacement model cannot be found among already discovered structures. Therefore, more laborious experimental methods involving collecting additional data sets are typically required. A method has been developed for predicting protein structures that can use sequence and structural features of proteins less than 30% identical to the protein whose structure is being predicted. We present preliminary results showing that this structure prediction method can be used for generation of models of sufficient quality to serve as molecular replacement models. Thus we hope to reduce the time and expense required to solve protein crystal structures.

---

***Designing Consensus-degenerate Hybrid Oligonucleotide PCR Primers for Identifying Novel Genes and Unknown Pathogens Using the New iCODEHOP System***

**Presenter:** Richard Boyce

**Author(s):** Richard Boyce, Tim Rose, Tobias Mann, A. Gregory Bruce, Jonathan Ryan, Jeanette Staheli

**Abstract:** We have developed a novel PCR technology that rapidly and specifically identifies distantly related orthologs and paralogs belonging to a common gene family. The method utilizes pools of PCR primers called CODEHOPs (Consensus-Degenerate Hybrid Oligonucleotide Primers) derived from highly conserved regions of multiply aligned protein sequences from members of a gene family. The CODEHOP technology has proven invaluable for the identification of unknown viral pathogens as well as the identification and structural characterization of viral and cellular genes implicated in various biological processes and disease. We have developed software to design CODEHOP primers and made it freely accessible on the WWW. The software, called iCODEHOP, is an interactive, Open Source, Web application. In addition to simplifying degenerate primer design workflows, iCODEHOP extends and improves the CODEHOP method of designing degenerate PCR primers from protein multiple alignments. The new iCODEHOP program implements an improved version of the original CODEHOP

algorithm and provides interactive visualization of PCR primers designed using the algorithm. Users can quickly scan over an entire set of degenerate primers produced by the program to assess their relative quality and select individual degenerate primers for further analysis. The program predicts annealing temperatures for degenerate primer pools, displays phylogenetic information for the sequences covered by the primer, and allows the user to easily design new degenerate primers from subselections of their input sequences. In this presentation, we will discuss the iCODEHOP program and demonstrate the design of CODEHOP primers for the identification of unknown pathogenic adenovirus species.

---

***A New Attempt to Stimulus Related Data Analysis by Structured Neural Networks***

**Presenter:** Bernd Brückner

**Author(s):** Bernd Brückner, Tobias Walter

**Abstract:** In the analysis of biological data artificial neural networks are a useful alternative to conventional statistical methods. A combination of a Liquid State Machine (LSM) and the Multilevel Hypermap Architecture (MHA) is used for analysis of stimulus related data, exemplified by fMRI studies with auditory stimuli. The MHA is an extension of the Hypermap introduced by Kohonen. Results from investigations with this structured neural network show an improvement of discrimination in comparison to statistical methods. With an interface to the well known BrainVoyager software and with a GUI for MATLAB an easy usability of the MHA and a good visualization of the results is possible. The combination of a Liquid State Machine with MHA as a readout, called LSM-MHA, is a new attempt for connecting the different time scales of spiking models with self-organizing maps. The data analysis with this kind of structured networks offers new possibilities for stimulus related data analysis.

---

***Evaluating Coevolution Detection Methods on Protein Alpha Helices***

**Presenter:** J Gregory Caporaso

**Author(s):** J Gregory Caporaso, Lawrence Hunter, Rob Knight

**Abstract:** Correlated evolution (coevolution) between positions in biological sequences is a source of important information about biomolecules including structural, allosteric, and intermolecular interactions. Coevolutionary data is however, difficult to acquire due to the stochastic nature of sequence evolution, varying degrees of relatedness between sequences in input alignments, the multiple-comparisons problem involved in analyzing all pairs of positions in alignments, and in proteins in particular, the lack of biologically relevant evaluation data. We propose using alignments of alpha helical proteins to address the lack of evaluation data for judging coevolution detection methods. Ionic interactions between stacked residues in alpha helices are known to be important for alpha helix stability, and we therefore expect these positions to coevolve. We applied a variety of automated methods for detecting coevolution to alignments of alpha helices, and observe coevolutionary signal between positions 3, 4, and 7 residues apart, coinciding with the periodicity of the alpha helix. Alpha helices therefore provide a biologically relevant data set for evaluating and comparing high-throughput methods for

detecting protein coevolution.

---

***Mobile Software for Blood Glucose Monitoring and Management***

**Presenter:** Yoojin Chung

**Author(s):** Yoojin Chung, Yoojin Chung

**Abstract:** It is very important to monitor blood glucose data of diabetics regularly. We use mobile and web technologies so that the burden of measuring, monitoring, and management of data can be reduced. A cellular phone that is equipped with a glucose-meter measures and transmits a diabetic's glucose data to a diabetics-care data server. The cellular phone also receives medical treatment or dosage information. In this paper, we describe the implemented software for cellular phones of the mobile diabetics-care system. In user interface design, the main menu consists of Measure, Analysis, and Manage. In Measure, the window guides a glucose tester to insert a test strip and to test blood glucose maximum three times. Measured data is stored in each diabetic data area and sent to a health-care server. In Analysis menu, the values such as amplification ratio, gain, testing time and voltage can be adjusted for the calculation of a glucose data with the measured raw data. The Manage menu manages accumulated data and shows the detailed analysis data for each diabetic. From the menu specification, we drew use case diagrams for menu behaviors and we extracted classes for overall software system. We implemented the classes based on WIPI, one of the integrated mobile standard platforms. The implemented software consumes only 1,160 KB memory footprint at maximum. Compared with other systems, our mobile equipment supports not only server-based data processing but also local data processing. And each cellular phone can handle multiple diabetics data.

---

***Analysis of Highly Conserved Functional Sites Predicted by Dynamics Perturbation Analysis (DPA)***

**Presenter:** Judith Cohn

**Author(s):** Judith Cohn, Michael Wall, Dengming Ming

**Abstract:** We previously developed an algorithm called Dynamics Perturbation Analysis (DPA) to predict functional sites in protein structures (D. Ming, M.E. Wall. 2006. J Mol Biol 358:213). The high throughput version of this algorithm (Fast DPA) decorates the surface of a protein with test points and uses approximate calculations to characterize the degree to which each point perturbs the protein's thermal vibrations. Residues near points that cause a large change are predicted to reside in functional sites. We applied Fast DPA to over 50,000 domains from SCOP. From this analysis, we predicted over 63,000 functional sites. We were able to obtain sequence conservation data from the HSSP database for 60,466 of the predicted sites. From these we selected 5,020 with high sequence conservation. In most of these highly conserved sites, we were able to associate the predicted residues with either an inferred ligand-binding site (determined by proximity to a small molecule in PDB), a catalytic site (from the Catalytic Site Atlas) or both. However, our methods did not yield annotations for 433 of the predicted sites

(corresponding to 132 SCOP families). Among these families, we identified 118 in which transitive annotation might be possible using annotations for similarly located DPA clusters in other members of the family. In 14 other cases, no obvious transitive annotation was possible. We present a more detailed analysis of the DPA predictions for these 14 families.

---

***Profile Fuzzy Hidden Markov Models for Phylogenetic Analysis and Protein Classification***

**Presenter:** Chrysa Collyda

**Author(s):** Chrysa Collyda, Sotiris Diplaris, Pericles Mitkas, Nicos Maglaveras, Costas Pappas

**Abstract:** This paper proposes a method for aligning multiple genomic or proteomic sequences using a fuzzyfied Hidden Markov Model (HMM). HMMs are known to provide compelling performance among multiple sequence alignment (MSA) algorithms, yet their stochastic nature does not help them cope with the existing dependence among the sequence elements. Fuzzy HMMs are a recently introduced type of HMMs based on fuzzy sets and fuzzy integrals which generalizes the classical stochastic HMM, by relaxing its independence assumptions. In this paper, the fuzzy HMM model for MSA is applied in HPV virus protein homologies. Comparative experiments with the classical HMM and other well-known approaches depict that fuzzy HMMs can increase the model capability of aligning multiple sequences mainly in terms of average identity percentage and computation time. In phylogenetic analysis multiple sequence alignments are obtained from the model and are used for the phylogenetic analysis of viruses coming for the HPV family. The results of the analysis are compared against those obtained by the classical profile HMM model and depict the superiority of the fuzzy profile HMM in this field. In protein classification, profiles are rescored using the fuzzy HMM approach. This process leads to a different representation of protein data in terms of motifs, since the new scores obtained by the definition of the fuzzy HMM allow for the discovery of novel motifs that may exist in protein sequences.

---

***Assembling the Interactome of Human Extracellular Matrix to Understand its Role in Health and Disease***

**Presenter:** Graham Cromar

**Author(s):** Graham Cromar, John Parkinson

**Abstract:** The extracellular matrix (ECM) is a self-assembling, fibrous network of proteins secreted by cells. The assembled matrix acts as a composite material and plays a role in a number of biologically important processes including cell adhesion, migration, proliferation and differentiation providing orientation during the development and migration of tissues as well as structural support. There is strong evidence that changes in these functions are associated with diseases as apparently diverse as arthritis, atherosclerosis and cancer. Despite the availability of powerful new tools and approaches for exploring higher levels of organization in such systems very little attention has been

devoted to structural proteins and the role of network connectivity in their self-organization and biomechanical properties. To address this problem, we constructed an initial map of the human ECM interactome consisting of 361 nodes and 547 edges from literature curated interaction data hosted at BioGRID ([www.thebiogrid.org](http://www.thebiogrid.org)). We found that annotations of human ECM proteins deposited in the Gene Ontology are largely incomplete as compared with their corresponding mouse or rat orthologues. A comparative analysis of these orthologues and a review of GO terms used in our search has recently allowed us to expand our initial network 20-fold. In fact, we estimate that upwards of 2500 proteins may be involved in the organization of the extracellular space which approaches 10% of the human genome. This milestone will allow us identify functionally important network components involved in the evolution of multicellularity and aid our understanding of the ECM's role in health and disease.

---

***Consistency Assessment Among Multiple Probe Sets Interrogating the Same Gene on the Affymetrix MOE430 GeneChip***

**Presenter:** Xiangqin Cui

**Author(s):** Xiangqin Cui

**Abstract:** Affymetrix expression microarrays typically contain redundant probe sets that hybridize to different regions of the same gene for thousands for genes. The current standard statistical analysis methods only focus individual probe sets for differential expression. They rarely address how these individual probe sets differ in their responses to the treatment or condition under investigation. Even if cases where redundant probe sets respond differently are identified, it is difficult to interpret due to flawed probe set - to-target gene annotations. We describe a simple genome-based screening and grouping procedure that allows unambiguous assignment of redundant probe sets to groups in which each group member interrogates different regions of the same transcriptional unit. We compared results obtained from differential expression analysis using the standard Affymetrix annotations and the new redundant probe set groupings. We found that the new groupings substantially improve concordance among redundant probe sets. We also show that an additional filtering step based on Present-Absent calls from the MAS5 algorithm further improves concordance among redundant probe sets. The remaining non-concordant probe sets can be explained by potentially alternative transcript detected by different probe sets and degradation difference related to large distances between the probe sets. Based on this analysis, we conclude that the redundant probe sets need to be comprehensively considered to make an overall conclusion about the expression of the associated genes in response to the treatments/conditions.

---

***The RAST Servers -- Public Resources for high Quality Genome and Metagenome Annotation***

**Presenter:** Robert A. Edwards

**Author(s):** Robert A. Edwards, Daniel A. Paarmann, Folker M. Meyer, Ross Overbeek, Rick Stevens

**Abstract:** With thousands of microbial genomes and metagenomes being sequenced each year, high-quality, freely available, annotation pipelines are required to ease the burden of annotation. Using the subsystems-based annotation developed in the SEED we have created a highly efficient, open annotation pipeline for genome annotations. This turns the conventional annotations on their head, by annotating genomes based on their phylogenetic neighborhood.

A parallel pipeline was constructed to handle metagenomes, including those generated by pyrosequencing. These automatic bioinformatics systems are available to all at <http://www.nmpdr.org/>.

---

***Thermonucleotideblast: A Tool for Improved Assay-Dependent Searching of Nucleic Acid Sequence Databases***

**Presenter:** Jason Gans

**Author(s):** Jason Gans

**Abstract:** Nucleic acid-based detection assays are essential tools for identifying bacterial, viral and fungal pathogens in plants, animals and the environment. Successful assays must be both sensitive and specific, uniquely identifying small amounts of pathogen DNA or RNA in samples that contain large amounts of nucleic acid sequence from background species. In silico assay screening can reduce assay development costs by identifying both potential false positive matches and the absence of expected matches (false negatives). While there are a number of existing software programs for assay-dependent database searching, they are limited in choice of assay format, target database size and sensitivity of the search algorithm. In addition, most of the existing tools for in silico screening rely on heuristic definitions of sequence similarity based on either the number of mismatches (i.e. non-Watson and Crick base pairs) or the number of mismatches and number of gaps (insertions and deletions in the primer-template duplexes). We present a new computer program, ThermonucleotideBLAST, for screening in silico detection assays using the physically relevant criteria of query/target hybridization temperature and free energy change on binding. We will demonstrate that these physics-based search criteria offer improved receiver operator characteristics over existing heuristics-based algorithms, albeit with a higher computational cost. To offset this increased cost, ThermonucleotideBLAST has been parallelized using a novel adaptive query and database segmentation scheme that significantly reduces the required search time.

---

***Gold in the Slag Heap: Identifying Constraints on Horizontal Gene Transfer***

**Presenter:** Jason Gans

**Author(s):** Jason Gans, John Dunbar

**Abstract:** Horizontal gene transfer (HGT) is the acquisition of genetic information by an individual organism from a source other than its parent. Natural HGT between bacterial species plays a central role in bacterial evolution and has significant impact on economics, agriculture, and public health. Directed HGT (the intentional introduction of foreign genes into an organism) is a fundamental tool for both basic research and industrial biotechnology. Examples include DNA sequencing, protein production and the creation of transgenic plants and animals. HGT can give species access to beneficial genes and biochemical pathways that evolved in other lineages, and plays an important role in both the natural adaptation of pathogens (e.g. antibiotic resistance) and the maliciously directed evolution of new pathogens (e.g. addition of virulence genes). Whole genome shotgun sequencing projects are horizontal gene transfer experiments on a massive scale. We hypothesize that the expression of foreign genes and the overproduction of foreign DNA sequences carrying regulatory binding motifs in *E. coli* (the host for cloning) creates negative biases leading to over and under-representation of genome segments in sequencing libraries. Existing data from bacterial genome sequencing projects can be used to obtain a detailed model of the physiological constraints on HGT. Our preliminary results support our hypothesis and demonstrate our ability to model the impact of foreign DNA on *E. coli* and obtain novel insights into limitations of both natural and directed gene transfer in bacteria.

---

***Evolution of Trypanosoma Brucei Gambiense and T. b. Rhodesiense Based on Expressed Sequence Tags***

**Presenter:** Sery Gonedelé Bi

**Author(s):** Sery Gonedelé Bi

**Abstract:** African trypanosomiasis is caused by protozoan parasites, trypanosomes, which are transmitted by tsetse flies (of the genus *Glossina*). The disease occurs in two forms: a chronic form caused by *Trypanosoma brucei gambiense*, which occurs in West and Central Africa; and an acute form, caused by *T. b. rhodesiense*, which occurs in Eastern and Southern Africa. The chronic infection lasts for years, whilst the acute disease may last for only weeks before death occurs, if treatment is not administered. In the recent year, the amount of biological information being output by researchers has grown exponentially. More than 2000 African trypanosome expressed sequence tags (ESTs) have been sequenced. Expressed sequences are the genes, or protein coding portions of DNA, that are active within a cell. These sequences are valuable because they can be used for both gene discovery as well as to provide information on which genes are turned on in a cell at each stage of its life cycle. Many new *T. brucei* genes have been discovered in the past two years, some have homologues in other organisms while others are likely to be unique. Sequence comparisons of genomes or expressed sequence tags (ESTs) from related organisms provide insight into functional conservation and

diversification. Based on the public ESTs database, we will compare the sequence of EST of known gene from *Trypanosoma brucei gambiense* and *T. b. Rhodesiense*. Evolutionary mechanism underlying these genes evolution relative to trypanosomes infectiosity will be highlighted.

---

### ***Knowledge Integration for Gene Target Selection***

**Presenter:** Graciela Gonzalez

**Author(s):** Graciela Gonzalez, Juan Uribe

**Abstract:** Over the years, many methods have been proposed for facilitating the discovery of gene targets that underlie the pathology of different diseases. Most of these methods focus on the analysis and ranking (feature selection) of lists of genes and their expressions levels resulting from high-throughput experiments. There has been increased interest in using alternate data sources, such as Gene Ontology mappings or protein interaction data, for the same purpose. We present GeneRanker, an online system that allows researchers to obtain a ranked list of genes potentially related to a specific disease or biological process by combining gene-disease (or gene-biological process) associations with protein-protein interactions extracted from the literature, using computational analysis of the protein network topology to more accurately rank the predicted associations. In an automatic test to determine the validity of an association predicted by GeneRanker using co-occurrences in the literature for the gene and disease terms (using not only the number of publications but the ratio of occurrence of the two terms together versus the gene alone), the top 50 pairings reach precision levels of 96% (and 86% for the top 200), proving better than text extraction methods alone. The first steps of empirical validation were also taken, with the top 300 and a set of 300 low-ranked entries were queried against a whole-genome expression microarray for clinically-annotated glioblastoma tumors. The findings were generally consistent with the literature-based evaluations. GeneRanker is freely available online at <http://www.generanker.org>.

---

### ***Shape Contexts for Protein Binding Site Recognition***

**Presenter:** Concettina Guerra

**Author(s):** Concettina Guerra, Paola Bertolazzi, Sandro Cuzzolin, Alberto Paoluzzi

**Abstract:** It has been long recognized that shape plays an important role in molecular function. Thus several geometric techniques have been developed to compare protein conformations with the goal of understanding protein structure-function relationships. In this paper we apply computer vision techniques based on local descriptors to the problem of comparing two protein surfaces. Specifically, we propose the use of Shape Contexts to address the following three problems: 1) given two protein surfaces determine the regions on the two surfaces that are most similar; 2) given a complete protein surface and a template binding site find the region of the surface most similar to the template; 3) compare and classify binding sites. Shape contexts describe the regional shape of a three-dimensional object at a surface point P using the distribution of points in a support

region surrounding P. The support region, typically a sphere centered at P, is discretized into bins corresponding to a partition of the sphere into sectors and shells. Finally, an histogram is obtained by counting the number of points falling within each bin. Our preliminary experiments on comparing surfaces described in terms of shape contexts for several pairs of proteins binding to different ligands including ATP and NAD show that our method generally outperforms existing shape-based methods in terms of accuracy. The computation time however tends to be higher.

---

### *Classification Analysis of HIV RNase H Bioassay*

**Presenter:** Lianyi Han

**Author(s):** Lianyi Han, Yanli Wang, Steve Bryant

**Abstract:** In this work, we present a new chemical fingerprint probabilistic neural network (CFPNN) to classify the compounds according to their bioactivities in HIV-1 reverse transcriptase associated ribonuclease H (RT-RNH) assay. The neural network is constructed by using the chemical fingerprint representation as input and the probabilistic evaluation of the classification as output through the Gaussian-shaped kernel. The 10-fold Cross Validation is used for the early stop of the network training and the hold-out method is employed for the extra model validation. Our results suggest that CFPNN has the potential to learn from a large collection of compounds by correlating their chemical fingerprints to HIV-1 RT-RNH inhibition activities, thus to classify the diversified compounds into the bioactivity categories. Prominent fingerprints were selected based on the Information theory and analyzed for understanding on the explicit connections between bioactivity and molecular fingerprints. The proposed CFPNN model can be potentially further employed in validation and noise reduction of given bioassay, as well as to be used in virtual screening for hits exploration.

---

### *A Support Vector Machine Method to Classify Enzyme Modulators*

**Presenter:** Katherine Herbert

**Author(s):** Katherine Herbert, Virginia L. Iuorno, Jeffrey H. Toney

**Abstract:** Our study investigates a support vector machine (SVM) method for the classification of enzyme modulators. The application of predictive classification to structure-activity relationships is a tremendously valuable cheminformatics approach for the discovery of potential therapeutic agents. The data analyzed was obtained from publicly available high-throughput screening libraries of small molecules in comma-separated value (CSV) and structure data format (SDF) files. Bioassay and stereochemical data were ranked according to compound properties, including substance ID, activity (inactive, decreasing, increasing), EC50 (half-maximal activity), Hill coefficient (plot slope), number of chiral centers, molecular substructure, bond type, and individual atoms. The simplified molecular input line entry system (SMILES) enabled comparison of substructure text strings and provided the crucial and key link between the target and test compounds. The supervised machine learning model was implemented using molecular feature vectors with associated labels through the SVMlite software.

Target data was trained to delineate a decision boundary separating molecular feature sets for classification and creation of the SVM learning model. The SVM model was subsequently applied to the test data to label compounds as inactive or active. Preliminary results were obtained for pyruvate kinase, glucocerebrosidase, and phospholipase A2. Three-fold cross validation of the model was performed on a training set with known labels and yielded 97.6% prediction accuracy. Compounds predicted to be active, i.e., inhibitors or activators, were found to cluster for particular structural groups that are known to exhibit therapeutic effects.

---

### ***Universal Geometrical Code in Proteomics***

**Presenter:** Sriram Kannan

**Author(s):** Sriram Kannan

**Abstract:** Physical Geometric algorithms generate infinite data sets that could be used for further data analysis leading to new insight into structural proteomics field. In such an attempt, different proteins belonging to different SCOP classes were selected and their amino acid sequences were downloaded from NCBI site and were also subjected to analysis of isoelectric point. Based on this two data sets an internal data set is generated based on MS EXCEL programmed algorithm that does the following data processing automatically when the protein sequence is uploaded from an internal database in EXCEL. 1. The linear amino acid sequence is converted to a scattered points data set on the basis of the isoelectric value of each amino acid residue. 2. Then virtual lines are drawn from the first and last residue location of each amino acid residue to each residue location of Tyrosine, Serine and Threonine intersecting perpendicularly on the virtual line created at the protein's isoelectric point in view of residues near phosphorylation sites and giving importance to protein function. 3. The distance is estimated for each virtual line automatically by this excel based algorithm and infinite datasets are generated that have the potential of further analysis. (The formula used is the distance formula used in mathematical sciences for calculating distance between two points). 4. In such an analysis it was found that one data set exhibiting the distance between inter amino acid residues calculated for all residues exhibits statistically significant data set that is discussed in the poster using the Excel program generated graphics. Thanking you Sriram

---

### ***Genetic Algorithms Select Protein Features Most Predictive of Enzyme Function***

**Presenter:** Andrew Kernytsky

**Author(s):** Andrew Kernytsky, Burkhard Rost

**Abstract:** Often we know the function of a protein and the residues responsible for this function. More often, we know neither and face a situation wherein the relevant local sequence features elude discovery. The challenge then becomes to sample a vast space of possible features without over-fitting the experimental observations. We introduced a methodology combining genetic algorithms with neural networks and support vector machines. We demonstrated the power of our concept by applying it to predict whether or not a protein is an enzyme and if it is, which type of enzyme it is without using direct

homology-based inference. The success of the genetic algorithm originated from its capability to sieve through a vast space of intersection features and to zoom into very specific aspects of enzymatic activity. The method thereby discovered very local motifs from global data.

---

***Semi-supervised Learning for Protein Sequence Classification***

**Presenter:** Brian King

**Author(s):** Brian King, Chittibabu Guda

**Abstract:** Protein sequence data continues to become available at an exponential rate. Annotation of functional and structural attributes of these data lags far behind, with only a small fraction of the data understood and labeled by experimental methods. This represents an ideal domain to explore semi-supervised learning methods, which are employed in situations where there is both labeled and unlabeled data available, and the amount of unlabeled data often excessively exceeds the amount of labeled data available. Classification methods that are based on semi-supervised learning have been shown to increase the overall accuracy of classifying unlabeled or partly labeled data in many domains, but very few methods exist that have shown their effect on protein sequence classification. We consider how proven methods in semi-supervised learning can be applied to classification of protein sequence data. We use the problem of protein subcellular localization prediction as the platform to demonstrate our work. Our approach is founded on application of the expectation-maximization algorithm applied to a Bayesian classifier. We consider both existing and novel extensions applied in this work, and demonstrate restrictions and differences that must be considered. Our results show that large repositories of unlabeled protein sequence data can indeed be used to improve predictive performance, particularly in situations where there are fewer labeled protein sequences available, and/or the data is highly unbalanced in nature.

---

***BANNER: A Portable, Open-source, Biomedical Named Entity Recognition System***

**Presenter:** Robert Leaman

**Author(s):** Robert Leaman, Graciela Gonzalez

**Abstract:** The increasing amount of research into biomedical named entity recognition has resulted in significant progress and created a need for a freely-available, open source system implementing the advances described in the literature. To fill this need, we have created BANNER, a trainable machine-learning system employing conditional random fields for biomedical named entity recognition (NER). It is intended to serve as a benchmark for gene mention recognition and as a study in portability of the techniques to other semantic classes of named entities. BANNER is designed to maximize domain independence by not employing semantic features or rule-based processing steps, and achieves significantly better performance than existing freely-available systems. It is therefore useful to developers as an extensible NER implementation, to researchers as a standard for comparing innovative techniques, and to biologists requiring the ability to

find novel entity mentions in large amounts of text. BANNER available for download at <http://banner.sourceforge.net>.

---

### ***Accuracy of Structure-based Sequence Alignment of Automatic Methods***

**Presenter:** Byungkook Lee

**Author(s):** Byungkook Lee, Changhoon Kim

**Abstract:** Accurate sequence alignments are essential for homology searches and for building three-dimensional structural models of proteins. Since structure is better conserved than sequence, structure alignments have been used to guide sequence alignments and are commonly used as the gold standard for sequence alignment evaluation. Nonetheless, as far as we know, there is no report of a systematic evaluation of pairwise structure alignment programs in terms of the sequence alignment accuracy. In this study, we evaluate CE, DaliLite, FAST, LOCK2, MATRAS, SHEBA and VAST in terms of the accuracy of the sequence alignments they produce, using sequence alignments from NCBI's human-curated Conserved Domain database (CDD) as the standard of truth. When the sequence similarity is low, structure-based methods produce better sequence alignments than by using sequence similarities alone. However, current structure-based methods still mis-align 11-19% of the conserved core residues, on average, when compared to the human-curated CDD alignments. The fraction of correctly aligned residues depends on sequence similarity and varies greatly among different protein pairs. Different methods performed differently for different structural types. DaliLite showed the most agreement with CDD on average, but performed relatively poorly for protein pairs that did not belong to one of the four major SCOP classes (all-?, all-?, ?/?, and ?+?).

---

### ***Detecting Reassortment in Influenza Viruses***

**Presenter:** James McInerney

**Author(s):** James McInerney, Victoria Svinti Svinti, James Cotton

**Abstract:** The Influenza virus has been a cause of worldwide infections in the human population for approximately 2,500 years, having a dynamic history characterized by common seasonal epidemics and occasional pandemics. The evolution of the virus during and in between these outbreaks is difficult to describe because it undergoes rapid evolution in order to evade the constantly adapting immune response of hosts. Its genome consists of eight individual segments of single stranded, negative sense RNA, each containing a single gene. In recent years, particular importance was given to the avian strains of flu due to its high pathogenicity, widespread infection of birds, rising number of cases identified in humans and its rapid distribution. The segmented nature of the genome allows for the exchange of entire genes between different viral strains when they co-infect the same cell through the process of reassortment. Identification of new reassortants is a crucial step in understanding viral evolution and in working towards preventing infections and the spread of fatal viruses. We have developed a method of

detecting true reassortments by a combination of SPR branch moves on phylogenetic trees derived from each segment and the use of a maximum likelihood test for the significance of difference in tree topologies. This test can identify real reassortment events and distinguish between real reassortment events and simple limitations of phylogenetic analysis.

---

***Automated Analysis of Viral Integration Sites in Gene Therapy Research Using the SeqMap Web Resource***

**Presenter:** Sean Mooney

**Author(s):** Sean Mooney, Jessica Dantzer, Brandon Peters, Sara Dirscherl, Scott Cross, Xiaoman Li, Kenneth Cornetta, Mary Dinauer

**Abstract:** Gene therapy holds great promise toward treating genetic disease, and many researchers are moving toward clinical application of this approach. One of the challenges faced is determining how viral integration events cause phenotypic changes in cells that eventually lead to cancer risk or malignant transformation<sup>1</sup>. Many recent studies have begun to address this challenge<sup>2,3</sup>. These studies, and their resulting publications, have described the abundance and location of insertional mutagenesis in normal and malignant cells using experimental methods such as ligation mediated (LM)-PCR or linear amplification-mediated (LAM)-PCR in both human and model organisms<sup>2,4,5</sup>. However, bioinformatic protocols are frequently used to describe genomic sites, and different gene annotations and different genome builds are used. Although the integration sites are fixed, these differences in genome builds and gene models can lead to differences, resulting in statistical analyses that are difficult to compare. To help researchers analyze these sites and the functions of nearby genes, we have developed SeqMap (<http://seqmap.compbio.iupui.edu/>), a secure, web-based, comprehensive integration site management tool that automatically analyzes and annotates large numbers of integration sites in mouse and human upon a common genomic framework. We believe use of this resource will enable better reproducibility and understanding of this important data.

---

***Molecular Evolution of Mammalian Imprinted Genes: Testing the Theories.***

**Presenter:** Mary O'Connell

**Author(s):** Mary O'Connell, Noeleen Loughran, Mark Donoghue, Karl Schmid, Charles Spillane

**Abstract:** At most autosomal loci the maternal and paternal copies of a gene are expressed in equal proportion. However, as discovered in the 1980s this rule is not steadfast. Major exceptions are found in placental organisms, where some autosomal genes are uniparentally expressed depending on the sex of their parent of origin. These are known as genomically imprinted (GI) genes. Currently for mammals there are ~100 genes of this type known. Many of these genes are crucial for the growth and survival of the offspring of a placental species. Disruption of these imprinted loci can result in a number of different disorders including Angelman and Prader-Willi syndromes, Alzheimers disease and autism. A number of evolutionary theories have been proposed

relating to the evolution of these GI genes, the most highly cited being the parental conflict theory, the gene dosage, and the ovarian time bomb hypotheses. However, support from molecular data for any one of these theories has not yet been elucidated. In this study we have used a computational and phylogenetic approach to determine the mechanisms of evolution of the known mammalian imprinted genes. Our results show that although these genes have evolved in a complex manner, their mode of evolution fits best with a gene dosage dependent model of evolution.

---

***ClearTK: A Framework for Statistical Biomedical Language Processing***

**Presenter:** Philip Ogren

**Author(s):** Philip Ogren, Philipp Wetzler

**Abstract:** The emergence of shared architecture and components for Biomedical Natural Language Processing (BioNLP) has made it much easier to leverage best-of-breed approaches for system development. I present ClearTK, a framework that facilitates feature extraction, training data creation, model building, and classification in the context of Statistical BioNLP. ClearTK decouples the task of feature extraction from the related tasks of model learning and instance classification, such that extracted features can be fed into one of several machine learning libraries for training and/or classification. By doing this, ClearTK makes it easier to experiment with different machine learning libraries using the same feature extraction library. ClearTK currently supports support vector machines (via LIBSVM), conditional random fields (via Mallet), maximum entropy (via OpenNLP), and several others that are provided in the WEKA library. In addition to providing a common interface for several machine learning libraries, ClearTK provides a rich set of feature extractors useful for BioNLP that can be mixed and matched in powerful ways, making it easy to experiment with many different feature sets using various machine learning libraries for a variety of BioNLP tasks. ClearTK is available under an academic use license and is built on top of the widely adopted Unstructured Information Management Architecture (UIMA).

---

***Gene Sequence Analysis of a Newly Isolated Lassa Virus Strain***

**Presenter:** Lawrence Okoror

**Author(s):** Lawrence Okoror, Fredrick Esumeh, Hillary Alaiya, Omatere Saiku

**Abstract:** Lassa virus is the cause of lassa fever which has been a leading cause of morbidity and mortality in many parts of West Africa. Recently, two new strains of the virus were isolated from Ekpoma Nigeria. Gene analysis was carried out with strain "Nig04-02". There have been no reason adduced for these yearly outbreak of lassa virus infection despite high level of circulating antibodies in the population. Sequence alignment of the gene predicted significant hits of between 85% to 45% with matrix set at PAM50, PAM100 and blosum 62 using blast and wu-blast tools. FGENESH, GENEID, GENSCAN was used to predict gene function Global alignment was done with wu-blast. Multiple sequence alignment was carried out using clustal W using both high scoring sequences and low scoring sequences. These were from mopeia and filovirus.

Transcription factors was predicted with TFsearch. And the likely gene product predicted with amigo. ORF was carried out with translate tool at expasy. Mathematical calculations were used to calculate the rate of deletions, insertions and substitutions. Sequence alignment (global) gave significant hits at PAM50, PAM100 and BLOSUM62 matrices which were between 85% to 45% homology. FGENESH gave no relevant prediction, while GENSCAN predicted 43.48% C G and Isochore 2 (43-51 C G%) and GENEID predicted a CDS1, CDSf, TSS, CDSi, PolA regions. All predictions were done at probability of >95%. There were 40 transcription factors with a score of 100% and six open reading frames. The rate of mutation was calculated, as well as the

---

### ***Metabolic Analysis of Tumor Progression***

**Presenter:** Norma Pawley

**Author(s):** Norma Pawley, Munehiro Teshima, Susan Carpenter, Steven Brumby, Clifford Unkefer, Pat Unkefer, James Freyer

**Abstract:** We present high-resolution metabolic profiles obtained from an in vitro tumor model using multi-cellular tumor spheroids. Despite the importance of metabolism to the development, progression and treatment of tumors, there are few studies of the cancer metabolome, per se. Essentially all work in the field of tumor metabolism has focused on studies of individual pathways or specific enzymes. Here we present metabolic profiles obtained from mass spectrometry analysis of multi-cellular tumor spheroids, and compare these profiles to those obtained from tumorigenic and non-tumorigenic cells, grown in standard culture and harvested either in exponential phase, or after the culture reached stationary phase (plateau). Results show significant differences between metabolites extracted from the four states, including a shift in glycolytic phenotype. Metabolic fingerprints are constructed for each state, and are used to interpret observed differences in metabolic profiles from small vs. large tumor spheroids.

---

### ***Reference Normalization Improves the Accuracy of Copy Number Analysis***

**Presenter:** Stanley Pounds

**Author(s):** Stanley Pounds, Cheng Cheng

**Abstract:** Most normalization methods for microarray data attempt to match parameters of the signal distribution across arrays. These methods assume that most of the discrepancies in these parameters in the initial signals are due to technical variation and error. However, in using SNP microarrays to study DNA copy number in cancer this assumption is not true. For example, hyperdiploidy is common in pediatric acute lymphoblastic leukemia (ALL) and thus it is intrinsically unreasonable to attempt to match parameters of the entire signal distribution across all arrays (especially with germline DNA samples). Thus, to study copy number alterations in cancer, the goal of normalization should instead be to match the signal distribution of unaltered genomic regions across arrays. We have developed a reference normalization algorithm for this purpose. Given a set of reference chromosomes for each study sample, reference normalization empirically defines array-specific transformations of the signal data that

maps the signal distribution for reference chromosome probe sets to a common distribution. From a theoretical perspective, this should greatly improve the accuracy of copy number analysis. These improvements are observed in practice in the study of ALL (Mullighan et al., Nature 2007) and other cancers. Reference chromosomes may be selected by the user with external data (such as cytogenetics) or computationally selected using features of the initial signal distribution. A computational reference selection algorithm successfully identified chromosomes with no cytogenetically evident abnormalities in 239 of 242 ALL samples.

---

***In vitro Functional Analysis of E. coli Photolyase Guided by Computational Coevolutionary Inferences***

**Presenter:** Steven Reuland

**Author(s):** Steven Reuland

**Abstract:** The enzyme photolyase is present in all three domains of life and likely represents among the first enzymes that evolved to repair UV damaged DNA. This enzyme uses energy captured from blue light to reduce cyclobutane pyrimidine dimers (CBDs) caused by UV exposure. Close relatives of photolyase have also evolved several times into blue-light regulated transcription factors in both plants and animals. To understand the functional significance of evolutionary change in this enzyme, we conducted extensive analyses to identify amino acids within the protein that appeared to undergo coevolution, or highly correlated evolutionary change. To experimentally test the inferred functional significance of these coevolving residues, we have studied E. coli photolyase in vitro. We have expressed, purified, and assayed several mutants of photolyase that represent coevolutionary intermediates or mutants with replacements at coevolutionary sites. Our experimental in vitro analyses support computational inferences and are consistent with coevolving sites representing groups of residues that have highly functional roles in the DNA binding activity of photolyase. These results help validate coevolutionary analyses as a means of effectively identifying residues that are likely to be functionally important.

---

***Combining Structural Modeling, Evolutionary Information, and Machine Learning to Improve Prediction of Nucleic Acid Binding Sites in Telomerase***

**Presenter:** Deepak Reyon

**Author(s):** Deepak Reyon, Ben Lewis, Jae-Hyung Lee, Colin Gleeson, Michael Hamilton, Fadi Towfic, Cornelia Caregea, Michael Terribilini, Drena Dobbs, Vasant Honavar

**Abstract:** Telomerase is a ribonucleoprotein enzyme that adds telomeric DNA repeat sequences to the ends of linear chromosomes. The enzyme plays pivotal roles in cellular senescence and aging, and because it provides a telomere maintenance mechanism for ~90% of human cancers, it is a promising target for cancer therapy. Despite its importance, a high-resolution structure of the telomerase enzyme has been elusive, although a crystal structure of an N-terminal domain (TEN) of the telomerase reverse

transcriptase subunit (TERT) from *Tetrahymena* has been reported. In this study, we explored the potential conservation of structural and functional features of TERT in phylogenetically diverse species. To identify amino acid residues in telomerase likely to make direct contact with either DNA or RNA, we developed improved classifiers for predicting nucleic acid binding sites in proteins. In addition to protein sequences, PSSMs and structural information were used as input for Na<sup>2</sup> Bayes and SVM classifiers. Structural models of the human and yeast TEN domains were generated by homology modeling and threading. In the context of these structures, comparison of predicted nucleic acid binding residues with available experimental results revealed significant similarities in the nucleic acid binding surfaces of *Tetrahymena* and human TEN domains. In addition, the combined evidence from machine learning and structural modeling identified several specific amino acids that we propose are important for binding, but for which no experimental evidence is available at present. Our results suggest that computational approaches can provide valuable information to guide experimental investigations of "recalcitrant" enzyme complexes.

---

### ***High-throughput Annotation of Genomic Datasets with Genephony***

**Presenter:** Alberto Riva

**Author(s):** Alberto Riva, Angelo Nuzzo

**Abstract:** Research in the biomedical fields increasingly requires the ability to manipulate and interpret very large datasets, due to the widespread adoption of high-throughput experimental methods and to the exponential increase of the amount of knowledge stored in public databases. This poses practical challenges that are beyond the skills of the average researcher: even conceptually simple tasks like integrating information from two different sources can become impractical if the datasets are very large, if the databases use different identifiers, and if (as is often the case) the user is not familiar with their internal structure and is not well-versed in database programming. Here we describe the initial implementation of Genephony, an online tool that facilitates the creation and manipulation of very large genomic datasets. Genephony allows the user to easily define sets containing biological entities (e.g. genes, SNPs, pathways, microarray probesets, etc), starting from files containing suitable identifiers, or from genomic regions. New sets can then be generated by deriving them from existing ones, or by combining two existing ones, and can then be browsed or exported in a variety of common formats. Relying on an extensive underlying database of genomic information, the system makes it easy to integrate, annotate and interpret the results of high-throughput experiments, providing automated operations that would be otherwise impractical if performed manually. Our expectation is that Genephony will become a useful tool for translational research, high-throughput biology, and for all knowledge-intensive data manipulation tasks in computational biology.

---

***Computational Analysis of Outer Membrane Proteins from Mycobacterium***

**Presenter:** Reatha Sandie

**Author(s):** Reatha Sandie, Michael Niederweis, Miguel Andrade-Navarro

**Abstract:** Outer membrane proteins (OMPs) play a role in virulence and drug resistance in mycobacteria, making them an important aspect of any functional Mycobacterium tuberculosis study. However, in contrast to more than 60 OMPs found in E. coli, to date only 2 have been found in mycobacteria. In a multi-step approach, we predicted OMPs in M. tuberculosis H37Rv and several other mycobacteria. OMPs were predicted using a series of physical and structural characteristics, based on a set of 30 gram-negative reference OMPs. Several families of potential transmembrane proteins were identified in the final list and show a high probability of being OMPs. Further analysis of the selected genomes allowed us to determine whether predicted OMPs are conserved among mycobacteria.

---

***Expression Profile of Human PMS2-Like Genes: Bioinformatic and Experimental Studies (cancelled)***

**Presenter:** Elena Shematorova

**Author(s):** Elena Shematorova, Dmitry Shpakovski, George Shpakovski

---

***Molecular Evolution of Human PMS2 Gene Family and its Possible Biological Implications (cancelled)***

**Presenter:** George Shpakovski

**Author(s):** George Shpakovski, Dmitry G. Shpakovski, Elena K. Shematorova

---

***SNP Analysis Reveals a Role for Polymorphisms in Evolution of Staphylococcal Pathogenicity***

**Presenter:** Karthikeyan Sivaraman

**Author(s):** Karthikeyan Sivaraman, Alexander Cole

**Abstract:** Single nucleotide polymorphisms (SNPs) represent small yet measurable steps in the evolution of genes and genomes in many bacterial species tested. In this work, we systematically analyzed the SNP content of the Staphylococcal minimal gene complement using 14 available genomes of Staphylococcus aureus. In the minimal coding region comprising 2001 genes, we found that 1917 genes (95.8%) had SNPs. Majority of SNPs were synonymous with only a fraction (

---

***Upstream Molecular Signatures of Constitutively expressed Genes in P. Falciparum***

**Presenter:** Geoffrey Siwo

**Author(s):** Geoffrey Siwo, Patrick Duffy, Moses Limo

**Abstract:** The mechanisms involved in the regulation of *P. falciparum* genes still remain elusive. The genome encodes only a few transcription associated proteins as compared to other genomes of the same size. In addition, only a few regulatory motifs have been identified. The regulation of genes in this parasite, however, occurs in a highly timed and stage specific manner. In this study, we used machine learning methods to determine whether 139 constitutively expressed (expressed in 4 parasite stages) and 136 stage specific genes could be discerned using the frequency of di- or tetranucleotides in their upstream regions. We applied K-means and expectation maximization algorithms to determine whether the natural partitioning of genes into groups corresponded to their categorization as constitutively expressed or stage specific. We also applied a Na? Bayes classifier to determine whether these upstream molecular signatures could be used to predict constitutively expressed and stage specific genes. The results showed that 65.7% of constitutively expressed genes cluster into a distinct group using upstream tetranucleotides while 72.8% cluster into a single group using upstream dinucleotide frequency. The Na? Bayes classifier, based on the tetranucleotide frequency attained a specificity of 77.85% and 76.2% based on dinucleotide frequency at a sensitivity of 77.8% and 75.95% respectively, by leave-one-out cross-validation. These results indicate that even without knowledge of regulatory motifs, the expression of genes in *P. falciparum* can be modeled using oligonucleotide frequencies in the upstream regions.

---

***Dimension Reduction in Association Study between Polymorphisms and the Response to Montelukast in Asthma (cancelled)***

**Presenter:** Jiun Su

**Author(s):** Jiun Su, Hsing-Kuo Pao

---

***Protein Structural Domain Assignment with a Delaunay Tessellation Derived Lattice***

**Presenter:** Todd Taylor

**Author(s):** Todd Taylor, Iosif Vaisman

**Abstract:** We describe a fully automated method of protein structural domain assignment using a Potts model which we call DePot (an abbreviation for Delaunay-Potts). Each amino acid residue is represented as a site in an irregular lattice derived from the Delaunay tessellation of the protein structure. Domain membership is represented by a spin value and each site has a spin which can change under the influence of its neighbors. Spins are allowed to interact subject to an Ising ferromagnetic-like energy function until clusters of like spins emerge and these clusters define domains. DePot is simple and easy to implement and the assignments agree well with previously published methods as we show on a large test set.

---

*Statistics of the Geometry and Topology of Real and Model Protein Structures*

**Presenter:** Todd Taylor

**Author(s):** Todd Taylor, Iosif Vaisman

**Abstract:** We have subjected several sets of real and simplified model proteins to Delaunay tessellation and have computed statistics on both Delaunay simplex geometry and the tendency of quadruplets of residue types to be joined together in simplices. We have characterized the geometry and contact patterns of real proteins and some of the ways in which they differ from these model structures. We have also found heretofore unreported asymmetries in contact patterns among residue quadruplets joined in simplices in real proteins.

---

*Combining Interaction and Expression Data to Identify Context-Specific Gene Sub-Networks*

**Presenter:** Hannah Tipney

**Author(s):** Hannah Tipney, Sonia Leach, Weiguo Feng, Richard Spritz, Trevor Williams, Larry Hunter

**Abstract:** We present a methodology which combines expression data and interaction information to highlight functional associations which previously had only tenuous support from the interaction sources. We demonstrate our method in mouse which has few known interactions and complex biology, providing examples of context-specific networks found using craniofacial expression data.

---

*Minimax Estimation of Means with Applications to Microarray*

**Presenter:** Tiejun Tong

**Author(s):** Tiejun Tong, Liang Chen, Hongyu Zhao

**Abstract:** The development of microarray technology has revolutionized biomedical research, and microarrays have become a standard tool in biological studies. Due to the cost and experimental difficulties, it is common that thousands of genes are measured only with a small number of replicates. In particular, the standard gene-specific estimators for means and variances are unreliable and the corresponding tests usually have low power. Contrary to the recent development on improving the estimation of variances, little attention has been paid to improving the estimation of means. In this poster, I will introduce a new family of shrinkage estimators for the mean vector under the assumption of unequal variances. The proposed estimators are proved to be minimax under the quadratic loss function. We also conduct simulations to evaluate the performance of our proposed estimators and construct a shrinkage-based t-like statistic that utilizes information across genes. Both simulations and real studies show that the

shrinkage-based test provides a powerful and robust approach for detecting differentially expressed genes.

---

***Bayesian Networks for Genome Expression: A Bayesian Statistical Approach to Modeling Gene Regulatory Pathways in Human Placental Microarray Data***

**Presenter:** Elinor Velasquez

**Author(s):** Elinor Velasquez

**Abstract:** Bayesian statistical thinking is considered by many as a revolutionary force within genetics and bioinformatics. A Bayesian network is a graphical model that encodes statistical relationships among a set of variables. In this way it can be used as an exploratory in-silico biological tool to infer novel gene expression regulatory pathways in a given genome. Our hypothesis is to confirm known and predict new gene expression pathways in the human placenta. Human placental gene expression time series Affymetrix data is obtained from the Fisher laboratory, U.C. San Francisco. Using a 2 Gigabyte, 2 Gigahertz Mac computer, we employed a variety of learning algorithms, including K2, hill-climbing, and simulated annealing to create a final gene expression network of 40 gene expression profiles for human placenta. Our results confirmed and inferred a number of novel gene expression regulatory pathways in the human placenta. A portion of this network will be confirmed by molecular genetic techniques. To summarize, we performed exploratory analysis of gene regulatory pathways of the human placenta (maternal-fetal interface) during normal pregnancy using Bayesian network methodology. The results obtained are consistent with known regulatory pathways in the human placenta, and predict interesting connections not previously discerned by standard bioinformatic analyses.

---

***Improving Topology Prediction Accuracy by Adding Reentrant Regions to the Topological Grammar***

**Presenter:** Håkan Viklund

**Author(s):** Håkan Viklund, Arne Elofsson

**Abstract:** As alpha-helical transmembrane proteins constitute roughly 25% of a typical genome and are vital parts of many essential biological processes, structural knowledge of these proteins is necessary for increasing our understanding of such processes. Because structural knowledge of transmembrane proteins is difficult to attain experimentally, improved methods for prediction of structural features of these proteins is important. OCTOPUS, a new method for predicting transmembrane protein topology is presented and benchmarked using a dataset of 124 sequences with known structures. Using 10-fold cross validation, OCTOPUS predicts the correct topology for 92% of the sequences, compared to 85% for the second best method. In particular, OCTOPUS is the first method to fully integrate prediction of reentrant regions into topology prediction. This is shown to provide an increase in topology prediction accuracy of 6 percentage units.

---

***Phylogenetic Analysis of Protein Phosphatases of the Malaria Parasite Plasmodium Falciparum***

**Presenter:** Jonathan Wilkes

**Author(s):** Jonathan Wilkes, Christian Doerig

**Abstract:** We report an exhaustive analysis of the *P. falciparum* genomic database (PlasmoDB) aimed at identifying and classifying all protein phosphatases (PP) in this organism (the phosphatome). This survey is placed in an evolutionary context by comparison with the PPases detected in a diverse range of eukaryotic genomes. **RESULTS**(i) Within the highly conserved PPP family exactly one *P.falciparum* sequence was present in each sub-group. (ii) Malarial PPMs clustered into distinct clades, a behaviour seen in other species, especially *A.thaliana* which has undergone a massive expansion of this family of PPase (iii) Two potential dual-specificity phosphatases were found(iv) A single prenylated tyrosine phosphatase type IV was the only PTP detected in *P.falciparum*(v) The *P.falciparum* phosphatome has the fewest members of any of the organisms surveyed (cf *H.sapiens* 103, *A.thaliana* 141 sequences)**Conclusion:** The considerable phylogenetic distance between Apicomplexa and other Eukaryotes is reflected by profound divergences between the phosphatome of malaria parasites and that of mammalian cells. The minimal nature of the *Plasmodium* phosphatome indicates a potential lack of redundancy compared with host organisms suggesting this may be an appropriate target for chemotherapy.